Named Entity Identifier for Malayalam Using Linguistic Principles Employing Statistical Methods

Bindu.M.S[1] and Sumam Mary Idicula[2]

**[1] Dept.of Computer Science ,M.G University
Edappally, Cochin, Kerala, India**

**[2] Dept. of Computer Science, Cochin University of Science and Technology
Cochin, Kerala ,India**

Abstract

Natural language processing (NLP) began as a branch of Artificial Intelligence is a field of computer science and linguistics and is concerned with interaction between human language and computer. Major tasks of NLP such as Machine Translation (MT), Information Retrieval (IR) and Summarization require extensive knowledge of the language for the effective identification of semantic information in the text. Meaning or semantics of a text is mainly decided by the named entities which are the role carrying agents in a text. The system presented here is a Named Entity (NE) Identifier created using Statistical methods based on linguistic grammar principles. Malayalam NER is a difficult task as each word of named entity has no specific feature such as Capitalization feature in English. NERs in other languages are not suitable for Malayalam language since its morphology, syntax and lexical semantics is different from them. For testing this system, documents from well known Malayalam news papers and magazines containing passages from five different fields are selected. Experimental results show that the average precision recall and F-measure values are 85.52%, 86.32% and 85.61% respectively.

Keywords: *Malayalam compound word, Finite state Transducer, Extended Conditional Random Field, Feature vector***.**

1. Introduction

NER is an important tool in almost all natural language processing applications such as IE, IR and Question Answering (QA) systems. Proper identification and classification of NEs are very crucial and pose a big challenge to the NLP researchers. The level of ambiguity in NER makes it difficult to attain human performance [1].
NER is the process of identifying and categorizing names in text. The NE task was first introduced as part of the MUC 6 (MUC 1995) evaluation exercise and was continued in MUC 7(MUC 1998).This formulation of NE task defines 7 types of NE: PERSON, ORGANISATION, LOCATION, DATE, TIME, MONEY and PERCENTAGE. NER also known as entity identification and extraction is a subtask of IE that seeks to locate and classify atomic elements in text into predefined categories. In the expression named entity the word named restricts the task to those entities for which one or many rigid designators as defined by Kripke stands for the referent [2].

The term named entity is not strict but has to be explained in the context where it is used. Entity names form the main context of a document. NER is a very important step towards more intelligent IE and management. NER performs what is known as surface parsing ,delimiting sequences of tokens that answer important questions such as “what”, “where” and “how” in a sentence.

Malayalam belongs to the Dravidian family of languages and is one of the 4 major languages of this family. It is one of the 22 scheduled languages of India with official language status in the state of Kerala. It is spoken by 35.9 million people. Malayalam is a morphologically rich agglutinative language and relatively of free order. Also Malayalam has a productive morphology that allows the creation of complex words which are often highly ambiguous [3]. A lot of work has been done in the field of NER for English and European Languages. In English Capitalization is a major clue for identifying person names. Some efforts have been made for Telugu, Hindi and Bengali. As of now we have no information regarding Malayalam NER work and no tag set is been identified so far.

Conditional Random Field (CRF) is a probabilistic framework for labeling or segmenting data. It is a form of undirected graphical model in which each edge represent conditional dependencies between random variables at the nodes. Each random variable Y_i is conditioned on an input sequence X. The conditional dependency of the random variable on X is normally represented by some feature functions [4] [5]. This feature function varies according to the application. CRF is commonly used for the labeling of natural language text or biological sequences. They were first used for the task of shallow parsing by Lafterly et al (2001) where CRF were mainly applied for Noun Phrase (NP) chunking. In CRF, with respect to figure 4, Y is dependant only on X while high order CRF or Extended CRF represent a model in which each Y_i is dependants on X as well as on n number of previous variables Y_{i-n}, …., Y_{i-1}

Regular Expression is the standard notation for characterizing text sequences. Finite State Automata (FSA) is a mathematical device used for implementing texts represented by regular expression. A variation of FSA called a Finite State Transducer (FST) is a machine that reads a string and outputs another string. Formally an FST is represented by a 6-tuple [6]. FST's applications are in speech recognition, phrase chunking, POS tagging etc.

Most of the Question Answering systems require answers which are either nouns, adjectives, adverbs or phrases. Deriving these NEs from large collection of documents is a difficult task. Currently there are QA systems available in different languages where they are using keyword extraction techniques.

2. Related Works

NER is a process of finding mentions of specific things in running text. It is an essential tool for QA and IR. But research indicates that NER systems are brittle meaning that NER systems developed for one domain do not typically perform well on another domain. Various approaches available for solving such problems are statistical machine learning techniques, rule based systems and hybrid approaches.

Machine learning methods are using either supervised learning or unsupervised learning techniques. Statistical methods require large amount of manually annotated training data. Few commonly used statistical methods are Hidden Markov Model (HMM), Maximum Entropy Model (MEMM) and Conditional Random Field (CRF). Sequence labeling problem can be solved very efficiently with the help of HMM. The conditional probabilistic characteristics of CRF and MEMM are very useful for the development of NER systems. MEMM is having a POS label biasing problem. But all machine learning techniques require large relevant corpuses which is unavailable in Malayalam. Machine learning methods are cost effective and no need of much language expertise. In [7] authors describe a NER system using CRF. This system uses different contextual information of the words along with both language independent and language dependant features. Paper [8] proposes a HMM based on the mutual information independence assumption where they claimed that their system reaches 'near human' performance. NER system based on MEMM is presented in [9].

Grammar based techniques are used for creating NER systems that obtain better precision but at the cost of lower recall and months of work by experienced computational linguistics. Rule based approaches lack the ability of coping with the problems of robustness and portability. Each new source of text requires significant tweaking of rules to maintain optional performance and the maintenance cost is quite high. Rule based systems performs the best especially for specialized applications. [10] Introduces a rule based system that use handcrafted rules and this approach gave them better performance than the CRF method.

Hybrid Methods either use combinations of different machine learning methods or combinations of rule based and machine learning methods. [11] Presented a tool for the recognition of NE in Portuguese. It has two components-rule based components for recognition of number expressions and hybrid component for names.Lot of work has been reported in the field of NER for English and European languages where one of the main features used is capitalization which is not present in Malayalam Language.

3. Malayalam Named Entity Identifier

NER is the process of identifying and categorizing names in text. In the taxonomy of computational linguistics NER falls under the domain of IE which extracts specific kinds of information from documents as opposed to more general task of document management which extracts all of the information found in a document.

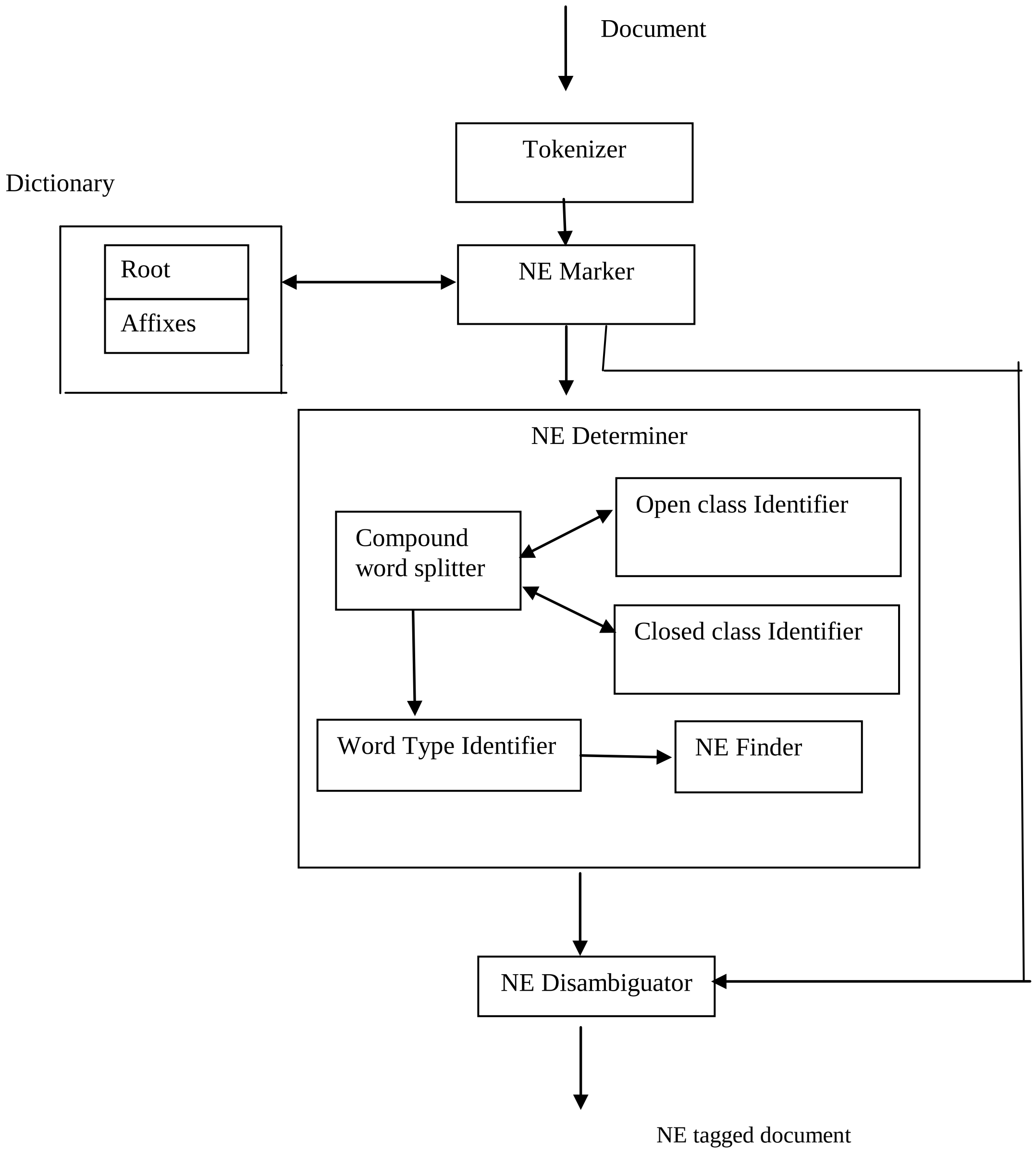

Figure 1: Named Entity Identifier

NE's are identified by using phonological, morphological, semantic, and syntactic properties of linguistic forms and that act as the targets of linguistic rules and operations. Two kinds of features that have been commonly used are internal and external, internal features are provided from within the sequence of words that constitute the entity-, in contrast, external features are those that can be obtained by the context in which entities appear [12].Based on the above investigation we have categorized an entity as either sole-entity, Constituent-entity, Dependant-entity or Not-an-entity.

3.1 Tokenizer

Input to the Tokenizer block in Fig 1 is a document in Malayalam. During the tokenization process each sentence of the document is taken and split into words or co-occurrence patterns.

3.2 NE Marker

This block checks each token to see whether it is present in the lexicon or not. Lexicon has all the root words along with its POS information. If it is present in the lexicon then it is a simple word, then the word details are retrieved from the lexicon. Based on this information token is marked with the possible NE tags.

Also, a word in the dictionary has information about their possible roles or named entities. But it is not possible to include all proper nouns in the dictionary. And another factor is that 85% of the words in Malayalam texts are compound words. Therefore obtaining NE tags from dictionary practically impossible.

3.3 NE Identifier

If the token M is a compound word then it is to be decomposed into its constituents M_1 to Mi.To find each constituent, the longest match method is adopted. When one component Mi is separated the remaining portion is sent to modification algorithm. The component M is searched in the lexicon, if it is not found transformation algorithm is called to obtain various forms of M and again searching is carried out. If not found process is repeated with next smaller string M. Based on the constituents, NE Identifier assigns suitable tags to each token [13].

Methodology - Finite State Transducer (FST)

Formally, a finite transducer T is a 6-tuple (Q, Σ, Γ, I, F, δ) such that: Q is a finite set, the set of states;
Σ is a finite set, called the input alphabet;
Γ is a finite set, called the output alphabet;
I is a subset of Q, the set of initial states;

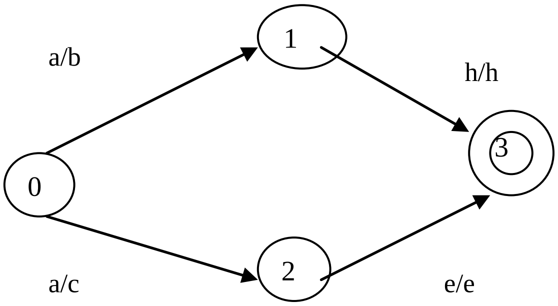

Fig 2 Finite State Transducer

F is a subset of Q, the set of final states; and δ is the transition relation.
Representation of the transducer in Fig.2 is
T=({0,1,2,3},{a,b,c,h,e},0,{3},{0,a,b,1},{0,a,c,2},
{1,h,h,3}, {2,e,e,3})
FST is a machine which accepts a string and translates it into another string. FST can also be used for generating and checking sequences [6]. A compound word is a string of Malayalam characters. To split this string into substrings an FST can be used.
Compound word splitter uses an FST with the following definition.
In the 6-tuple, set of states Q = {A,B,C,D,E,F,G,H}
Initial state I= {A}
Final states F= {C,D,E,F,G,H}
Input alphabet Σ = {compound words}
Output alphabet Γ = { valid simple Malayalam words}
Transition function δ= {NOUN,VERB,ADJECTIVE….,SUFFIX}
FST for compound word splitter is given in FIG.3

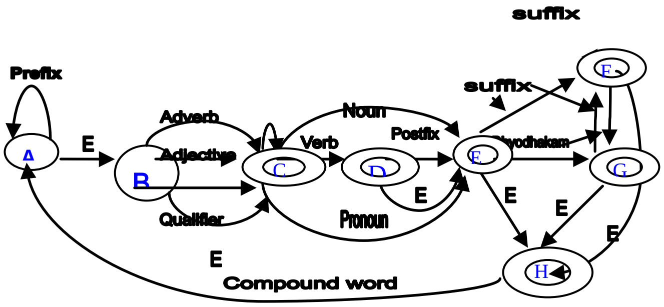

Fig.3 FST for Compound Word Splitter

This system operates in optimal time since the time to assign the tag to sentence corresponds to the time required to follow a single path in a deterministic finite state machine.

3.4 NE Tag Disambiguator

Previous blocks assigns each input token a single/multiple NE tags. Tokens with multiple tags are sent to the Disambiguator to solve the tag ambiguity which removes all tags except one. Output of tag Disambiguator is a string of all tokens along with their NE tags.

Methodology: Extended Conditional Random Field

Tag Disambiguator is implemented using high order CRF or extended CRF. It is an undirected graphical model in which each vertex represents a random variable whose probability distribution is to be inferred and each edge represents a dependency between two random variables. CRF's avoid the label bias problem, a weakness exhibited by MEMM. The primary advantage of CRF's over HMM is their conditional nature [4],

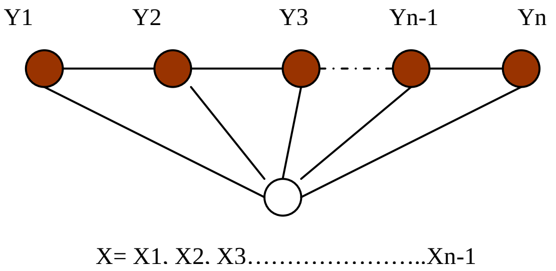

Fig4 Graphical structure of chain-structured CRFs

Let X={X1 …XN} and Y= {Y1…YN} be two sets of random fields. For the given input sequence X, Y represents a hidden state variable and CRF's define conditional probability distributions P (Y|X) over the input sequence. Sometimes the conditional dependency of each Yi on X will be defined through a fixed set of feature functions (potential functions) of the form f (i, Yi-1, Yi, X). The model assigns each feature a numerical weight and combines them to determine the probability of a certain value for Yi. CRF's can contain any number of feature functions and the feature function can inspect the entire input sequence X at any point during inference. CRF's are extended into high order models by making each Yi dependant on a fixed number of previous variables Yi-o ... Yi-1.

NE tagging can be modeled as a sequence labeling task where X= X1X2X3...Xn represents an input sequence of words and Y= Y1Y2Y3...Yn represents corresponding NE Label sequence. The general label sequence Y has the highest probability of occurrence for the word sequence X among all possible label sequences that is Y = argmax{Pr (Y|X)}. These labels are determined by the feature functions.

Main features for NE tagging have been identified based on the word combination and word context. The features also include prefix and suffix for all words [5].
Following are the features used for NE tagging in Malayalam.

- Constituents of current word: These determines the POS tag of the word as noun, verb etc.
- Context word features: Preceding (pw) and following words (nw) of the current word. We have taken pw1, pw2, pw3, nw1, nw2, nw3 as the feature.
- POS information: POS of previous words and in ambiguity resolution, POS of the following words are helpful.
- Contains digits or symbols. If the word contains digits they are marked with 'NUMBER' POS (cardinal number(CN) or ordinal number(ON))
- Lexicon feature: It contains Malayalam root words and their basic POS information such as noun, verb, adjective, adverb etc.
- Inflection lists: After analyzing various classes of words inflection lists of nouns, verbs and participles are prepared to improve the performance of the POS tagger.

Working of Named Entity Identifier

1. Tokenize the document.
2. Check each token whether it is present in the dictionary, if it is a simple word then retrieve the word details and determine its NE category
3. If not present in the dictionary and if it is a compound word, call a compound word splitter to obtain its constituents. Use this information to find out the NE type of the current word
4. If above two steps are not sufficient to determine the NE type, call NE detector using ECRF
5. Repeat steps 2-4 for all the tokens

   NE Marker determines whether a token is a sole entity or not.NE Determiner marks a token with constituent-entity tag and Tag Disambiguator with Dependant-Entity tag. All other tokens are labeled with Not-an-Entity tag.

4. Tests and Discussions

NER is designed and implemented using J2SDK1.4.2 and MySQL. Its performance is evaluated using standardized

Table 1: Performance of NE Identifier

Token Type	***NE/NAN***		
	Precision	***Recall***	***F-Measure***
Proper Noun	60.0	73.0	65.86
Pronoun	85.2	87.4	86.29
Common Noun	81.4	83.5	82.44
Locative	86.3	85.0	85.64
Accusative	87.6	88.1	87.85
Dative	84.1	87.2	85.62
Instrumental	89.3	88.3	88.79
Reason	81.0	82.4	81.69
Sociative	91.7	89.0	90.33
Car-Num	90.6	91.2	90.89
Ord-Num	93.1	92.3	92.69
Adj-Num	89.0	87.5	88.24
Adj-Quantity	86.0	87.3	86.64
Other Tokens	92.0	90.5	91.24

techniques precision, recall and F-score where Precision is defined as a ratio of number of correct NER tags to the number of NER tags in the output and recall is the ratio of number of correct NER tags to the number of NER tags in the test data. F-score = 2*recall*precision/ (recall+ precision) [14].

Documents related to five different fields are selected as the corpus. Then we randomly selected 8000 sentences for training and 2000 sentences as test set. Precision, recall and F-score obtained for various types of NE tags are shown in table 1. We could overcome the following challenges raised by the Malayalam language features by considering the word level and phrase level information ie by morphological analysis, POS tagging and phrase chunking.

Agglutinative Nature

Malayalam is a highly inflectional and agglutinative language. 85% of words in Malayalam text are compound words and hence role of these words can be decided only by knowing its components and their types. Role of an entity depends on the importance of the word which is decided by local and global information. To derive local information, each word is analyzed and collected its component details.

Word Order

Malayalam sentence is a sequence of words where words may appear in any order and each word can be a combination of any number of stems and affixes. Even though there is no specific order for the words in the sentence, within a chunk word categories are related.

In Malayalam language there is no distinction between uppercase and lowercase. Hence proper techniques are to be adopted to overcome such challenges.

5. Conclusion

Main task of NER is to identify and classify named entities in a given document. Identification is concerned with marking the presence of a word/phrase as NE in the given sentence and classification is for denoting the role of the identified NE.

Most of the systems have concentrated on three kinds of NEs ie on the roles of proper nouns, Time and percentage expressions. But these entity types are not sufficient for many question answering systems where entities like reason, cause, instrument etc are to be identified. The NER system described here is designed incorporating these types. Also this paper addresses the problem of NER in a query which involves the detection and classification of the named entity in a given query into predefined classes.

We have selected formal text since this is developed as a part of QA system based on health IR. For this application text from various textbooks, journals and magazines and web sites are selected which are mostly formal texts.

6. References

[1] Stefan Schwarzler Joachim Schenk,Frank Wallhoff and Gunther Ruske,"Natural Language Understanding by Combining Statistical methods and Extended Control Free Grammars",Proceedings of 30th DAGM Symposium on Pattern Recognition,Springer-Verlag Berlin,Heidelberg,2008.

[2] Lev Ratinov Dan Roth, "Design Challenges and Misconceptions in Named Entity Recognition", Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL), pages 147–155, Boulder, Colorado, June 2009

[3] A .R .Rajarajavarma,"Keralapanineeyam", National Book Stall, Kottayam, 2000.

[4] Hanna.M.Wallach, "Conditional Random Fields", University of Pennsylvania CIS Technical Report MS-CIS-04-21.

[5] Samir AbdelRahman, Mohamed Elarnaoty, Marwa Magdy and Aly Fahmy," Integrated Machine Learning Techniques for Arabic Named Entity Recognition", IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 4, No 3, July 2010

[6] Bindu.M.S, Sumam Mary Idicula,"Analysis of Malayalam compound words and Implementation of a compound word splitter tool using Finite State Models", International Conference on Modeling and Simulation India 1-3 December 2009.

[7] GuoDong Zhou Jian Su ," Named Entity Recognition using an HMM-based Chunk Tagger", "Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 473-480.

[8] Mohammad Hasanuzzaman1, Asif Ekbal2 and Sivaji Bandyopadhyay3," Maximum Entropy Approach for Named Entity Recognition in Bengali and Hindi", International Journal of Recent Trends in Engineering, Vol. 1,No.1, May 2009

[9] Burr Settles," Biomedical Named Entity Recognition Using Conditional Random Fields and Rich Feature Sets", Proceedings of the COLING 2004 International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA). Geneva, Switzerland. 2004.

[10] Gjorgji Madzarov, Dejan Gjorgjevikj and Ivan Chorbev," A Multi-class SVM Classifier Utilizing Binary Decision Tree", Informatica 33 (2009) 233-241

[11] B. Sasidhar, P. M. Yohan,Dr. A. Vinaya Babu, Dr. A. Govardhan," Named Entity Recognition in Telugu Language using Language Dependent Features and Rule based Approach", International Journal of Computer Applications (0975 – 8887) Volume 22– No.8, May 2011

[12] Xiaofeng Yu,"Chinese Named Entity Recognition with Cascaded Hybrid model",Proceedings of NAACL HLT 2007 Companion Volume, pp 197-200,April 2007

[13] Asif Ekbal and Sivaji Bandyopadhyay, "Named Entity Recognition Using Appropriate Unlabeled Data, Post-processing and Voting ",Informatica 34 (2010) 55–76

[14] Alireza Mansouri, Lilly Suriani Affendey, Ali Mamat, "Named Entity Recognition Approaches", IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.2, February 2008

Bindu.M.S received her B.Tech degree from M.A College of Engineering, Kothamangalam in 1986 and M.E degree from Coimbatore Institute of Technology, Coimbatore in 1988.She is currently pursuing the Ph. D. degree in the research area of Natural Language Processing from Cochin University of Science and Technology, Cochin, India.

During 1988-1998 she was with Manipal Institute of Technology, Manipal, as Lecturer and then as Reader in the Department of Computer Science and Engineering. Currently she is working as Reader in the Department of Computer Applications with Mahatma Gandhi University, Kottayam India. She has published several papers in International and National conference proceedings. Her research interests include Natural Language Processing, Artificial Intelligence and Information Retrieval.

Dr. Sumam Mary Idicula took B.Sc (Engg) degree in Electrical Engineering from College of Engineering Trivandrum in 1983. She pursued her Master studies in the field of Computer and Information Science in Cochin University of Science & Technology and took M.Tech degree in 1986. She started her carrier as lecturer in the Department of Computer Science of Cochin University of Science & Technology in 1987. She took PhD degree in Computer Science later and is now working as Reader in the same Department.

She is an active researcher in the field of Natural Language Processing and Human Computer Interaction. She has undertaken 3 major projects supported by ISRO and UGC in the field of Natural Language Processing and 2 major projects supported by AICTE and KSCSTE in the field of Human Computer Interaction. She is guiding several M.Tech students & Ph.D Scholars. About 40 research papers have been published by her in the field of Computer Science in reputed journals and in international conferences. She has visited Europe and United States for participating in International Conferences & Workshops.

She is a member of the Board of Studies of Computer Science and Board of Studies of Computer Applications of Cochin University of Science & Technology and also a member of the Academic Committee of CUSAT.